Predicting stressful working conditions of office workers using webcam and other unobtrusive sensors from their work computer*

Lefkothea Patrikiou[2734169]

VU Supervisor: Dr Michel Klein,
Daily Supervisor: Dr Peter-Paul van Maanen, **

Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam

Abstract. Stress is a common issue in the modern workplace, negatively impacting both employee welfare as well as organizational efficiency and productivity. Detecting stress early is crucial to prevent its long-term effects. However, current methods for detecting stress in naturalistic settings are limited. This study aims to address this gap by leveraging the sensors embedded in office workers' laptops to predict stress levels during office work. Through a controlled experiment that simulates the naturalistic setting of knowledge workers, and comparison of different classification approaches, we successfully distinguish between stressful and neutral working conditions with an accuracy of 58% using a Decision Tree model. Our findings show that the choice of data splitting strategy to training and test set greatly affects the model's performance. Moreover, we demonstrate that personalized models, tailored to each participant, achieve classification accuracy ranging from 58% to 100% for some participants, with most predictive features being the ones that describe head and gaze orientation. Collectively, our results underscore the efficacy of personalized approaches in predicting stressful working conditions during office work.

1 Introduction

Stress is a pervasive problem in today's fast-paced work environment, with significant implications for both employee well being and organizational productivity [1]. Office workers, in particular, are at an elevated risk of stress due to the constant demand for high mental workloads and the need to adapt to rapidly evolving technological landscapes [2]. Early detection of stress is crucial to prevent it from becoming chronic and causing irreparable harm. Unfortunately, an automatic, continuous, and unobtrusive method for detecting stress in naturalistic settings is currently lacking.

* Supported by Welpair Solutions B.V.
** Welpair B.V. Padualaan 8, 3584 CH Utrecht

The laptop used by office workers can be seen as a collection of sensors. We can hypothesize that they can be used to monitor behavior, well-being, and environmental conditions through the capture of video, audio, and typing behavior data. The advantage of using the sensors embedded in workers' laptops is that they are unobtrusive and one does not require additional measuring devices, making them easy to use "on the fly".

Welpair is a company dedicated to promoting the well-being of office workers and fostering safe and healthy working environments. The Welpair Assistant software aims to leverage the sensors embedded in workers' laptops to monitor, among others, their posture and working environment in real-time. Hence the aim of this thesis to objectively and accurately measure workplace stress levels from such sensors aligns well with the goals of Welpair.

Video data, from which one can track facial expressions, body language, and even physiological indicators such as heart rate, have the potential of being a particularly useful form of unobtrusive measurement given that most individuals have a webcam on their work computer.

The novelty of the overarching project created by Welpair lies in the integration of heart rate estimation from video data, with facial and typing behavioral measures obtained through unobtrusive sensors. This approach aims to enable real-time prediction of stress levels in office environments. Additionally, a separate sub-project will focus on posture classification. The posture modality can later be used as a feature in the stress classification as it has been shown to be a good indicator of assessing the stress levels of workers [3]. The focus of this current project is to address the task of binary stress classification during office work.

As such, we can pose the following Research Question : Can we utilize data from multiple sensors on the workers' laptop, such as video from webcam and typing behavior data from keyboard and mouse, to predict stressful working conditions during office work? To answer this, we break down our main research question into smaller sub-research questions, each addressing a distinct aspect:

1. Can we induce work-related stress in a controlled experimental setting?
2. Which Machine Learning (ML) approach is the most suitable for modeling work-related stressful conditions?
3. Which features provide the most useful information for accurately detecting stressful situations at work?

To tackle these questions we will conduct a literature review to identify relevant features used in similar tasks and identify the different machine learning algorithms for classification. Subsequently we will describe an experimental setup where stress levels induced in participants will be assessed through questionnaires, and data will be gathered to evaluate the different classification approaches. We will also investigate whether a generalized approach, where the model can effectively generalize to new users, is more advantageous compared to a personalized approach, where distinct models are tailored to different users. Finally, we will evaluate the feature importance of the best-performing model to

gain insights into which features offer the most valuable information for stress detection.

2 Literature Review

In this section we aim to identify related work in the literature as well as pinpoint features that have been proven useful for tasks similar to ours. Finally we identify and describe the learning algorithms for classification that are well-suited for our purpose.

2.1 Related Work

Over the last few years efforts have been made towards automating stress detection using machine learning models, with some studies focusing on stress specifically in office environments [4–6]. However, the majority of these efforts have relied on training models using physiological data obtained from body sensors, limiting their applicability for real-time stress estimation in office settings [7]. In this section, we will identify the most useful physiological features, also discussing whether they can be estimated from video data. We will also emphasise features that can be extracted from data gathered by computer sensors such as video and typing behavior.

Physiology

It has been shown that increased stress levels cause increase in Heart Rate (HR), blood pressure, Galvanic Skin Response as well as decrease in Heart Rate Variability (HRV), rendering them important stress markers [8–10]. Especially HRV is widely used in stress detection and based on the findings of Hjortskov et al. in [11] it was concluded that heart rate variability (HRV) serves as a sensitive indicator of mental stress in office environments and that it is a better predictor of stress than Blood Pressure, since it is regulated by "central command", whereas blood pressure is regulated in the periphery and is influenced by local factors and conditions. However our aim is to only use unobtrusive measures. Recently, there has been great effort towards accurately estimating HR measures from video [12]. Therefore we aim to estimate HRV using available methods [13, 25].

In [14] a real time stress detection system is presented, that monitors participants in a simulated office work setting over three days. They used ECG and acceleration sensors and got an average detection accuracy of 98%.

Furthermore, in a highly relevant work they detected stress in working environments using several work computer sensors, as well as some physiological features [3]. They found that an Support Vector Machine (SVM) model was able to differentiate between neutral and stressful working conditions with an accuracy of 90%. When looking into the best features for such a prediction, posture, followed by facial features were found to be the most valuable ones in

such a setting. A few physiology features were on the 3rd place (HR and skin conductance)

Video derived features

Gaze: Liao et al[15] developed a real-time stress detection system when working with a computer and showed that eyes focus on computer screen more frequently and for longer periods of time during stressful situations.

Posture: In their study, [18] analyzed the changes in posture of office workers by utilizing a device that measured the pressure in their chairs. Twenty-eight men were subjected to the Montreal Imaging Stress Task (MIST) [19] to induce stress, and it was confirmed that the amount of fast movement increased during stress tests as compared to control. Such a finding is highly relevant for us as movement can be easily measured with the webcam. Posture has also been analysed using visual techniques, such as skeletal tracking, for example in [20] they deducted that posture can be an indicator of the motivation of the worker.

Facial features: Dinges et al. [21] developed an Optical Computer Recognition (OCR) technique to identify facial expressions indicative of stress caused by workload. They used self-reports, salivary cortisol measures, and HR signals as the ground truth. Changes in the eyebrows, mouth, and lips (including asymmetry, which is particularly useful for stress recognition) were tracked using a 3D deformable mask , and Hidden Markov Model (HMM) was employed to identify patterns of facial stress. They achieved classification results between 75% and 88% in distinguishing between high and low stress levels, thus confirming the potential of measuring eyebrow and mouth movements for this purpose.

Typing Behavior

Typing behavior measures are less commonly used for stress recognition than physiological measurements in the current state of the art. The reason is that they have received limited study, resulting in lower overall accuracy compared to physiological methods for stress detection [30]. Nevertheless, such measurements are of significant importance in this current project due to their unobtrusive nature and ease of measurement. Moreover, several behavioral measurements show promising potential:

Keyboard and mouse dynamics Keystroke and mouse dynamics have been extensively studied in the field of security, specifically for authentication purposes [26, 27], as well as in emotion recognition research [28] according to Kolakowska et al. However, they have not been widely used for stress detection. Regarding mouse dynamics, there are different opinions on its usefulness for detecting stress and emotions. Nonetheless, some studies [29] have made significant progress by acknowledging the reliability of behavioral biometrics derived from keyboard and

mouse usage patterns to evaluate stress levels and extract personality traits. For instance in [32] they used keystroke dynamics and succeeded to recognize 15 emotional states using decision trees. In another study [31] they investigated the effect of stress on keystroke features. Participants were instructed to write an email following a mentally or physically stressful task. Decision Tree, SVM, kNN, adaBoost, and NN classifiers were tested. kNN detected cognitive stress with an accuracy of 75%. Their findings provide evidence that cognitive stress has the ability to influence typing patterns.

2.2 Learning Algorithms for Classification

In this section, we will outline the learning algorithms that can be used for our purpose. To determine the most effective modeling approach for our problem, we describe the following types of classifiers as we expect them to perform well based on their performance on similar tasks in the literature review:

- Support Vector Machines (SVM)
- k-Nearest Neighbors (kNN)
- Classification trees: Decision Tree (DT) and Random Forest (RF)
- Artificial Neural Networks: Multilayer Perceptron (MLP)

SVM Support Vector Machines are supervised learning models widely used for a variety of classification task, showcasing excellent performance in a diverse range of applications [31]. SVMs aim to find an optimal hyperplane that separates different classes by maximizing the margin between them. They are particularly effective with high-dimensional data and are capable of handling non-linear relationships through the utilization of kernel functions. SVMs work well in classification tasks where the patterns of the classes in the data are well-separated, leading to the need for a clear decision boundary. However, SVMs may be less suitable for large datasets or when the classes exhibit significant overlap.

kNN k-Nearest Neighbors is a non-parametric classification algorithm that assigns a data point to a class based on the majority vote of its k nearest neighbors in the feature space. It does not make explicit assumptions about the underlying data distribution. kNN is performing well in classification when the patterns in the data are locally clustered and instances of the same classes tend to be close together in the feature space.

DT Decision Trees are hierarchical models that split the data based on the various features to make predictions. They are comprised of a series of decision nodes that recursively split the data, leading to leaf nodes that represent the predicted outcome. Decision Trees can handle both categorical and numerical data, rendering them a versatile model for various classification tasks. One of their key advantages is their interpretability, as the decision rules can be easily

understood and visualized. However, Decision Trees are susceptible to overfitting, which means they may become overly complex and struggle to generalize well to unseen data.

RF Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. Each decision tree is built using a random subset of the training data and features.This approach enables them to handle effectively high-dimensional data, capture non-linear relationships, and address missing values. They are well-suited for classification tasks that require feature importance rankings and can handle noisy datasets. However, they may underperform with imbalanced datasets.

MLP MLP Classifier is a type of artificial neural network with multiple layers of interconnected nodes, also known as neurons. It has the ability to learn complex patterns and relationships within the data using forward and backward propagation techniques. An MLP classifier can handle non-linear relationships and capture intricate features, making it suitable for classification tasks when the patterns in the data are more complex. However, training an MLP Classifier requires large amounts of training data and can be computationally expensive.

We decided to test all the above algorithms since they have all demonstrated to be effective in similar tasks in the literature.

3 Methodology

In this section we describe an experimental setup in which we aim to create a realistic office work setting in a lab environment where participants will perform a knowledge work task. Our experimental procedure is heavily inspired by [23] and aims to validate and expand their findings. Our purpose here is twofold: first, to validate our ability to induce work stress, and second, to describe and develop prediction models.

3.1 Experimental Design and Procedure

To collect data, the participants will perform a knowledge work task, namely writing two 2000-word essays on different topics per condition. Topics will be counterbalanced between participants and conditions, to prevent learning effects of topics. To manipulate the condition, we chose a relevant stressor in a knowledge worker context, namely time pressure to finish a task before a deadline. This setting has been shown to create a realistic office scenario [23]. Each participant (within-subject design) will undergo the two conditions below in the following order:

– Neutral: the participant is instructed to work on the task as long as they need. After a maximum of 45 minutes, the participant is asked to stop.

– Stress: the participant is given a deadline of 45 minutes to finish the task and is told that they should present on of their essays to the experimenter. A countdown clock showing the remaining time is provided.
We find this an effective work stressor based on the fact that the average person types 4.1 words per second [24], combined with the fact that existing research has demonstrated its effectiveness in inducing stress in office settings [23].

At the beginning of the experiment, participants will complete a brief "pre" questionnaire providing information on their gender, age, nicotine and coffee consumption, and the time frame during which they last experienced any kind of stress. After each block, participants will fill in a short questionnaire where they will self reported their stress levels ("how stressed are you now?", scale 1-5). In between the blocks there will be an 8 minute "relaxation phase", which is typical for stress research. The total duration of the experiment will be 3 hours. A timeline of the experimental procedure can be seen in Figure 1.

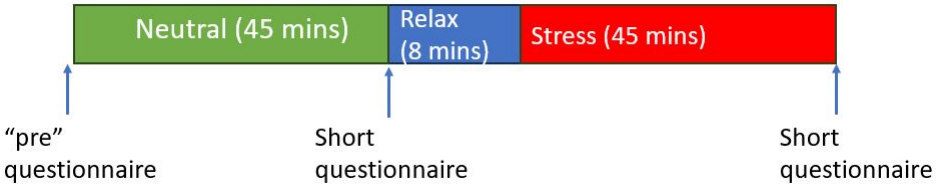

Fig. 1. Experimental timeline

Participants will be required to write essays on two different topics for each condition, consisting of one opinion topic and one information topic. We selected the topics from the following options:

– Opinion topics: Experience and opinion on work related stress/ on healthy living/ on internet privacy
– Information topics: Describe five tourist attractions in Perth, West Australia / Plan a coast-to-coast road trip in the USA / The life of Napoleon

Additionally, participants will be provided with instructions specifying the specific details to be included in their reports, such as statistics on the topic, positive and negative examples etc.

Participants will be allowed to look online for information, as well as use provided documents stored on the working laptop. This setting is representative of a knowledge work environment, as available information is combined with the worker's input to create new information [23].

To ensure that the participants will work on the tasks seriously, they will be told that the reimbursement they will receive will depend on the quantity and quality of the tasks they complete. To avoid bias, participants will be told that

the study is to investigate working patterns and wellness at the office, and they will be debriefed at the end of the experimental session.

Apparatus and Sensors for data collection Participants will perform the the tasks on a laptop (Acer Aspire 5 or HP 15s) with Windows 11. They will be provided with a keyboard, mouse and a monitor (Philips LCD monitor 27E1N5600AE). The apparatus will be placed on a table, at a standard distance from the edge of the table, and participants will be seated on standard office chairs. The essays will be written on Libre Office and the presentation will be prepared in google slides. The internet browser will be google chrome on guest mode.

To record video a web-camera will be placed on the monitor (Microsoft LifeCam, HD-3000). The height of the monitor and camera will be adjusted per participant so that their eyes are located on the top 1/3 of the screen. One video file of approximately 3 hours will be recorded for each participant. Videos will be captured at 30 fps and will be saved in mkv format using a custom made script developed in-house. Audio will also be independently recorded from the camera, used to timelock each condition.

Typing behavior will also be recorded with the same custom script made in-house. The script records:

- a timestamp and the keycode at every key pressed on keyboard
- at every mouse click: the timestamp, x,y coordinates, button id (left or right) and cumulative button hold time
- at every mouse move: the timestamp , x,y coordinates, cumulative distance moved as well as cumulative time moved
- when the mouse scrolled: the timestamp and cumulative scrolling time

These data will be stored as event-based SQL tables.

Participants Our study will be advertised via flyers in Vrije University, Utrecht University and the co-working space Utrecht Inc. It is aimed for university students and young professionals. Given their experience from their coursework and office work, and their use of computers as their primary tool, such participants are considered a representative sample of knowledge workers. Furthermore, they are experienced on our specific experimental knowledge work tasks; namely report writing and preparing presentations. They will receive a standard reimbursement for their participation.

The quality of their written essays and presentations will be checked in order to assess if they will be properly engaged during the task. If their level of engagement is deemed inadequate, they will be excluded from the study.

3.2 Data Extraction

Facial behavior We will first timelock and extract the video clips of neutral and stress conditions from the complete 3-hour video recordings. To accomplish

this the audio files will be utilised and manually scanned to identify auditory cues indicating the start and end of each condition. The timestamps corresponding to the audio duration (in seconds) when the cues are identified will be recorded and transformed into Unix timestamps (as the Unix time that audio starts recording + audio duration timestamp). Subsequently, the videos will be manually examined to determine the frame number corresponding to each timestamp. This video inspection process will be conducted using another custom, in-house-made script. Once we have compiled a list of all the start and end frame numbers corresponding to each condition for each participant, we will proceed to extract the video clips between these frames. A custom Python script employing the OpenCV module will be utilized for this purpose.

After extraction of the trimmed videos, facial behavioral analysis will be conducted for each participant. OpenFace, an open-source computer vision tool [33], will be utilized for this purpose. OpenFace is capable of detecting facial landmarks and head pose, recognizing facial action units (AU), and estimating eye-gaze in each frame of the videos. Facial AU are used to describe human facial expression. "Their intensity is measured on a range from 0 (not present) to 5 (present at maximum intensity) with continuous numbers inbetween" [34]. A comprehensive list of the relevant generated features and their descriptions can be found in Table 1.

Table 1. relevant OpenFace generated features

Feature name	Description
gaze_angle_x, gaze_angle_y	Eye gaze direction in radians in world coordinates averaged for both eyes
pose_Tx, pose_Ty, pose_Tz	coordinates of head location with respect to camera in millimeters. (positive Z is away from the camera)
AU01_r	Inner Brow Raiser Intensity
AU02_r	Outer Brow Raiser Intensity
AU04_r	Brow Lowerer Intensity
AU5_r	Upper Lip Raiser Intensity
AU06_r	Cheek Raiser Intensity
AU07_r	Lid Tightener Intensity
AU09_r	Nose Wringler Intensity
AU10_r	Upper Lip Raiser Intensity
AU12_r	LipCorner Puller Intensity
AU14_r	Dimpler Intensity
AU15_r	Lip Corner Depressor Intensity
AU17_r	Chin Raiser Intensity
AU20_r	Lip Stretcher Intensity
AU23_r	Lip Tightener Intensity
AU25_r	Lips Part Intensity
AYU6_r	Jaw Drop Intensity
AU45_r	Blink Intensity
AU45_c	Blink Presence
AU28_c	Lip Suck Presence

Typing behavior To extract the typing data corresponding to each condition, we will again utilize the Unix timestamps of the conditions. We will slice the SQL tables containing the typing data between the start and end timestamps of each condition for each participant. These slices of data for each condition will initially be stored in a dill format. Next, we will transform the cumulative event-based table slices, into a timeseries format. In this format, each event is represented as a separate column, eliminating the cumulative aspect. Specifically, we will convert the cumulative duration of key presses, mouse moves, mouse clicks, and mouse scrolls into the respective duration of each individual key press and mouse action (such as mouse move duration, click duration, and scroll duration respectively). We will accomplish this by calculating the differences between successive cumulative values for each measure. Additionally, we will transform the cumulative mouse move distance and cumulative mouse move duration into the duration and distance of each individual mouse move. Lastly, we will further transform the column containing the mouse button IDs (left or right) into two separate columns, representing the count of left clicks and right clicks, respectively. We will then use these extracted data to engineer some more typing behavior features, that are described in the respective paragraph.

3.3 Feature Engineering

Facial features The OpenFace output files will be further processed to derive additional"facial features" that have been flagged relevant from the related work. Specifically, the "gaze_angle" vectors will be used to determine the direction of the subject's gaze as follows: When both the "gaze_angle_x" and "gaze_angle_y" values are close to 0 (defined as the absolute values of both vectors being less than 0.1), it indicates that the subject's gaze is focused at the center of the screen in that specific frame. The "gaze_center" feature will be binary, with a value of 1 indicating the gaze was centered, and 0 otherwise. If the "gaze_angle_x" value is positive, it indicates that the subject is gazing to the left. In this case, the "gaze_left" feature captures the degree to which the subject's gaze is directed left. Similarly, when the "gaze_angle_x" value is negative, it indicates the subject's gaze is directed to the right. The "gaze_right" feature will represent the absolute value of the gaze vector, reflecting the degree to which the subject is looking right. Likewise, when the "gaze_angle_y" value is positive, it indicates the gaze is directed downwards, and that degree of downward gaze will be captured in the feature "gaze_down" . Similarly, when the "gaze_angle_y" value is negative, it indicates the gaze is directed upwards. In this case, the "gaze_up" feature will represent the absolute value of the gaze vector, reflecting the degree of upward gaze.

Additionally the feature "blink rate" defined as the number of blinks per minute was also calculated by counting the occurrences of blinks ("AUC45_c") in a minute while filtering out consecutive frames that were labelled as blinks within that minute.

Finally, facial data will be aggregated with a granularity of 1 minute for each participant. This choice of granularity is assumed to be appropriate consider-

ing the temporal dynamics of stress. Any type of stress typically takes time to manifest and unfold, making stress prediction at a sub-minute-level granularity unreliable. After all, the experience of stress is not characterized by rapid fluctuations where one moment we feel stressed and the next we do not. Finally this granularity was also used in [3] where they successfully distinguished between stressful and not stressful working conditions classifying each minute independently, which provided extra validity to our choice.

Typing features The timeseries data of typing behavior will undergo additional processing to derive useful typing behavioral features identified in the literature. Specifically, from the mouse usage data, we will calculate the "right_click_rate" and "left_click_rate" as the number of mouse clicks within a 60-second interval. The "mouse_speed" will be calculated as pixels/sec by dividing the distance moved by the duration of each mouse movement. Regarding the keyboard event timeseries, we will compute the typing speed in Words Per Minute (WPM), which is a widely recognized metric for quantifying typing speed. It is defined as the number of keys pressed divided by 5, approximating the number of words typed. This value is further divided by the total minutes of typing time. Additionally, we will calculate the "error_key_rate" which represents the frequency of pressing the "backspace" or "delete" buttons within a minute. Furthermore, the time spent not typing will be encoded by calculating the "typing_pause_rate," which indicates the percentage of time spent not typing within a minute.

The typing behavioral features will also be aggregated at a granularity of one minute. These aggregated features, along with the aggregated facial features, will be combined at the participant level. Finally, all participants' data will be merged to create our dataset. A list of all generated features can be found in table 2.

Table 2. List of all generated features

Facial, Head and Gaze Orientation Features		Typing features
AU1_InnerBrowRaiser	AU23_LipTightener	clickDuration
AU2_OuterBrowRaiser	AU25_LipsPart	RclickRate
AU4_BrowLowerer	AU26_JawDrop	LclickRate
AU5_UpperLipRaiser	AU45_BlinkInt	mouse_move_dist
AU6_CheekRaiser	AUc45_BlinkRate	mouse_move_duration
AU7_LidTightener	headOrient_x	mouse_move_speed
AU9_NoseWringler	headOrient_y	scroll_duration
AU10_UpperLipRaiser	headOrient_z	keyPressRate
AU12_LipCornerPuller	gazeCenter	pause_rate
AU14_Dimpler	gazeDown	errorKeyRate
AU15_LipCornerDepressonr	gazeUp	
AU17_ChinRaiser	gazeLeft	
AU20_LipStretcher	gazeRight	
AUc28_LipSuck		

3.4 Preprocessing

Once our complete dataset is compiled we will describe it in terms of instances and size and we will inspect it for missing values. In case of missing values, since most models do not handle them effectively, we will employ the Kalman filter [35] to fill them in. For this purpose we will utilise the pykalman python module. The Kalman filter offers the advantage of systematically filling in missing values and also detecting and correcting outliers.

The Kalman filter works by iteratively updating its estimate of the system's state based on new measurements. It holds two estimates: the predicted state estimate and the updated state estimate. The predicted state estimate is based on the previous estimate and the system's dynamic model, while the updated state estimate incorporates new measurements and corrects for errors. One of the key advantages of the Kalman filter is its ability to handle missing values and outliers in a systematic manner. When a value is missing, the Kalman filter can still provide an estimate by relying on the predicted state and the system's dynamic model. This allows for the reconstruction of missing values and the filling of gaps in the data [36].

Additionally we will examine the correlation matrix to observe the relationships among our features. This knowledge can aid in feature selection by ensuring that highly correlated features are not redundantly included in the models.

Finally, to asses whether our employed stressor leads to a typical behavior pattern, we will compare the means of each typing feature per condition. If we observe statistically significant differences in certain features between the conditions, we will select only those features to proceed with the rest of the analysis. For the comparison we will use the paired sample one sided t-test, as it is the appropriate statistical test for our analysis.

3.5 Classification implementation

In this section, we will describe the implementation of the classifiers that we chose to utilize. Furthermore, as we aim to validate and build up on the findings of the SWELL study [3], we will replicate and describe their process using our dataset. The key difference between their approach and our intended approach lies in the data splitting strategy employed to divide the data into training and test set, which we will elaborate on in the subsequent sections. The methodology introduced in this section will allow us to answer which ML approach is the most suitable for modeling work-related conditions.

Classification All models mentioned in the Literature Review section (kNN, RF, DT, SVM, and MLP classifier), will be implemented using the scikit-learn module (v 0.22.2) in Python 3. As a baseline benchmark we will use a simple majority baseline, which always predicts the majority class in the training data.

The hyperparameters of each model will be tuned using the random search method, using 100 iterations and will be evaluated on accuracy. The tuning process will involve performing a 5-fold cross-validation (CV) and will utilize the

Table 3. Search space of random search hyperparameter tuning

Model	Parameter	Values
SVM	c	21 values logarithmically spaced between 10^{-10} and 10^{10}
	gamma	21 values logarithmically spaced between 10^{-10} and 10^{10}
	kernel	rbf, poly
RF	number of estimators	100, 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000
	criterion	gini, entropy
	max depth	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, None
DT	max depth	2, 3, 5, 10, 20, 50, 100, 150, 200, 300, 500
	min samples leaf	5, 10, 20, 50, 100, 200, 500
	criterion	gini, entropy
kNN	n neighbors	range(1, 31)
	weights	uniform, distance
	metric	minkowski,euclidean,manhattan
MLP	hidden layer sizes	(10,30,10),(20,), (7,), (1,)
	activation function	tanh, relu
	solver	sgd, adam
	alpha	0.001, 0.001, 0.005, 0.05
	learning rate	constant, adaptive

relevant scikit-learn module. The search space for each model's hyperparameters can be found in Table 3.

The dataset will be split on a group-wise basis as our objective is that it can generalize over unseen participants. To perform the splitting we will employ scikits' GroupKfold method, with 80% (16 subjects) used as the training set and 20% (4 subjects) as the test set. Both sets will be independently standardised using the "StandardScaler" method of scikit-learn module. The training set will then be used for hyperparameter tuning, employing a group-wise splitting strategy within the cross-validation process. Model performance will be evaluated using accuracy as the metric.

The best-tuned model will subsequently be refitted using the full 80% of the training set and be evaluated on the test set, which will be comprised of participants that have not been seen before by the model. This evaluation aims to assess the model's generalization capabilities and its ability to perform on unseen data/users To ensure reproducibility, a random state of 42 will be used whenever randomness is involved. This includes the model initialization, the random search for hyperparameter tuning, as well as the data splitting process.

Comparison with SWELL: Splitting Strategy Once we have identified the best-performing model for our objective, we will proceed to compare its learning curve between the approached used in the SWELL study [3], and our current approach. In the SWELL study an accuracy of 90% is reported when distinguishing between stressful and non stressful working conditions, utilizing a random data splitting strategy for the train and test sets. In the random splitting, data is randomly allocated to 80%train and 20% test set, using the "train_test_split"

method from scikit-learn. Both sets will be independently standardised. The model will be trained using the training set and a 5-fold CV, in which data is also randomly split. Subsequently the model will be fitted on the entirety of the 80% training set and will be evaluated against the testing set, which contains never-seen-before participants' instances.

However as we aim for our model to be able to generalize over unseen participants we will also employ our group-wise splitting strategy, as described in the section above. So overall, in order to compare the effect of splitting strategy on the models learning and performance on the test set, we will begin with a dataset comprising of 7 participants (due to the 5-fold CV) and incrementally add one participant at a time to the dataset. For each dataset size we will apply the respective splitting strategies (random or group-wise), train the model using 5-fold CV (again utilising the respective splitting strategy in the CV), and assess its performance on the held-out test set.

Personalized Models To construct personalized models for each participant, the data from each individual will be randomly assigned to an 80% training set and a 20% test set. Both sets will be independently standardized. Subsequently, a 5-fold CV will be conducted using the training data to obtain the training accuracy. Next, the model will be fitted using the entire 80% of training data and will be applied to predict the held-out test set, thereby assessing its generalizability.

Feature Importance To determine the feature importance of each personalized model, we will utilize the "compute_feature_importance" method of decision trees from the scikit-learn module. To aggregate the feature importances across all personalized models, we will sum them and subsequently normalize the values to ensure they sum to 1. This will be achieved by dividing each individual feature importance score by the sum of all feature importance scores. This normalization allows for comparison of the relative importance of different features across different models.

4 Results

In this section we present the results of our experiment. The first subsection provides descriptive statistics of the dataset, while each subsequent subsection addresses one of the sub-research questions.

4.1 Participants, Preprocessing and Descriptive Statistics

Participants 24 university students and young professionals participated in our experiment of which 15 were female and 10 were male. Their ages ranged from 18 to 50 years old, with mean age of 21.36 and standard deviation of 8.46 years.

In the pre-questionnaire 4 participants indicated that they smoked a cigarette 30-60 minutes prior to the experiment. Furthermore 15 participants specified that they had experienced stress in the 2 hours prior the experiment. One participant did not conduct almost no work in the stress condition, was clearly not following the task guidelines and was hence excluded. Finally due to corrupt video files of 3 participants, our final dataset comprised of 20 participants.

Preprocessing and Descriptive Statistics We proceeded to analyze the complete dataset, which consisted of a total of 1200 instances distributed among 20 participants. Each participant contributed 30 rows of 1-minute aggregated data per condition, containing the facial and typing features described above. All the features in the dataset are numerical.

Regarding missing values, we observed that they were present in only certain typing features. Specifically, the columns related to mouse move distance, speed, and duration had 119 empty values each. The scroll duration feature had 528 missing values. Additionally, the typing features related to keypress duration and pause duration each had 60 missing values. We filled them in using the Kalman filter as described in the methodology section.

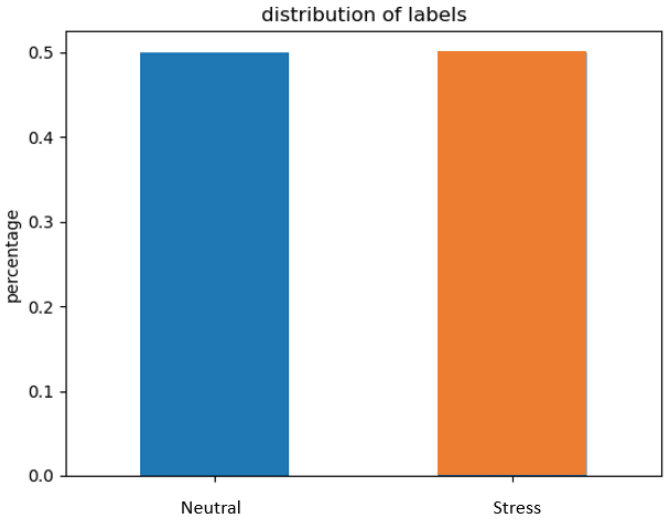

Fig. 2. Distrubution of labels in the dataset

Upon exploring the distribution of labels depicted in Figure 2, we saw that our dataset is balanced. This result is not surprising, as we allocated an equal amount of time for each condition during data collection.

Examining the correlation heatmap in Figure 3, we can observe that the features with the highest correlation coefficient are those that describe similar facial action unit (AU) activity. For instance, “AU12_LipCornerPuller” correlates

strongly (0.65) with "AU6_CheekRaiser" , which is understandable as pulling the corner of the lip upwards, raises the cheeks. Additionally, features that represent opposite directions of the same phenomenon also show relatively high correlation and should be treated with caution. For example, "gazeLeft" is negatively correlated (-0.45) with "gazeRight" and "head_orintation_x" correlates strongly (-0.98) with "head_orientation_y". Overall, the observed correlations among these features suggest that they are capturing related aspects of facial and head movements. This knowledge can aid in feature selection by ensuring that highly correlated features are not redundantly included in the models.

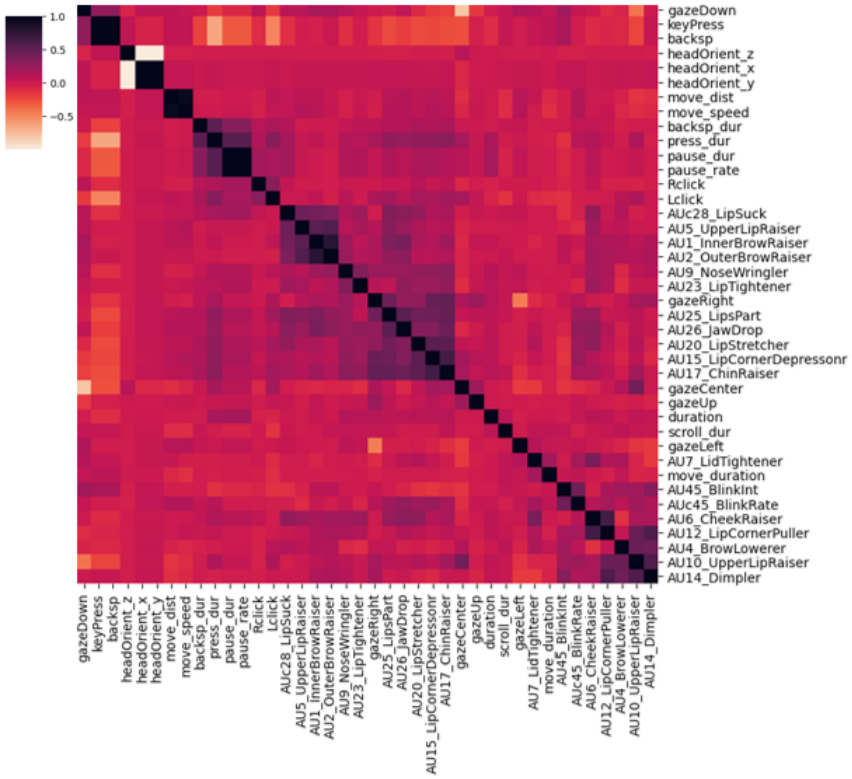

Fig. 3. Correlation matrix

4.2 Questionnaire results

In this section we present the questionnaire results, in order to asses if we successfully induced stress during our stress condition. We analyzed the questionnaire responses provided by the participants after the neutral and stress blocks.

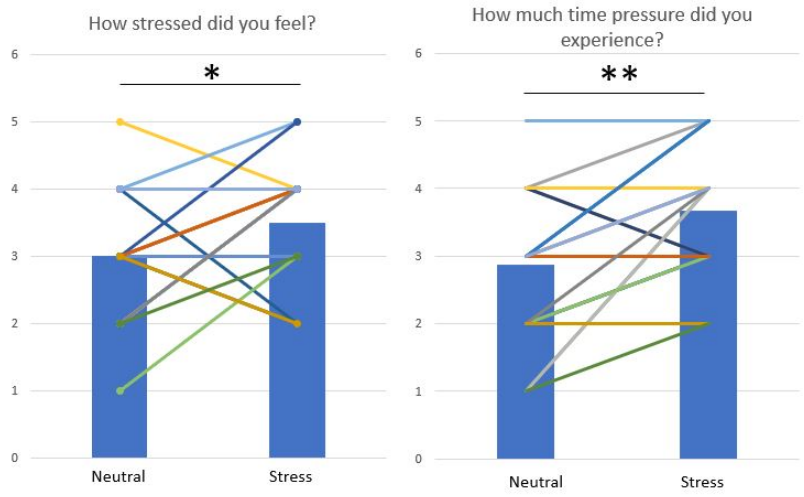

Fig. 4. Self-report scores of stress levels (left, p=0.0478) and time pressure (right, p= 0.0026) experienced after each block. Bars represent the mean, and lines represent the individual participants.

Figure 4 displays the mean self-report scores of stress levels and time pressure experienced in the neutral and stress conditions. The non-parametric Wilcoxon rank test, was used to compare the means of the rankings. This is the appropriate statistical approach for ranking data that are not normally distributed. The results of the self reported stress levels depicted on the left side of Figure 4, indicate a significant difference between the means of the two conditions, although the effect size is only marginal with a p-value of 0.0478 (t-statistic = 36) at a 95% confidence interval.

Regarding the time pressure experienced (Figure 4 right), the effect is more pronounced. A majority of participants (58%) reported experiencing more time pressure in the stress block compared to the baseline neutral block. The mean scores reported in the two conditions were significantly different as determined with the Wilcoxon rank test (p = 0.0026, t-statistic = 12).

To further assess if the stressor made an impact on the participants behavior we compared the typing behavior between the two conditions, by calculating the mean of each typing feature in each condition. One sided, within subjects t-test shows that there is statistically significant difference between the right click rate (t(19) = -3.195, p = 0.0045), left click rate(t(19) = -14.102, p = $7.47e^{-12}$), key press rate (t(19) = 2.272, p = 0.0342) as well as total mouse move duration (t(19) = -5.137, p = $5.02e^{-05}$) and total mouse move distance (t(19) = -4.241, p = 0.0004) between the two conditions, as depicted in Figure 5. So the stressor created a typical behavioral pattern.

The features which did not differ significantly between the two conditions can also be seen in Figure 5. As described in the Methodology section we decided to combine the facial features with only the statistically significant typing features in the models.

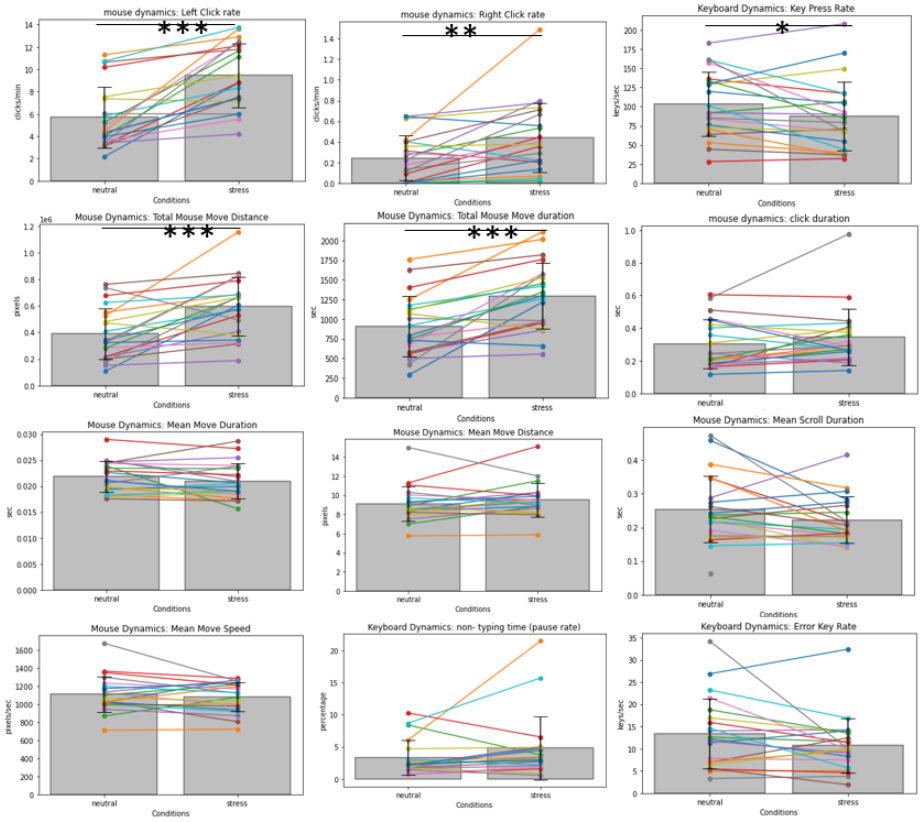

Fig. 5. Mean per condition of the different typing behavior features. Significant features are depicted with star between the bars. Bars represent the mean, and lines represent the individual participants.

4.3 Classification results

In this section we present the classification results of the different models we chose to use. Moreover we compare the splitting strategies and apply personalization on the best-performing model to determine the most suitable ML approach for predicting work-related stressful conditions.

We proceeded to train and tune the classifiers as described in the methodology section. All finetuned models were refitted on the whole of training data and

tested on the test set which contained never seen before participants. Classification results for predicting the working condition, as well as hyperparameters of the best model, are presented in Table 4.

Table 4. Classification Accuracy on training and testing set

Model	Train Accuracy	Test Accuracy	hyperparameters
Majority Baseline	-	0.5	-
SVM	0.63	0.47	rbf kernel, gamma: 1e-07, C: 10000.0
kNN	0.59	0.45	weights: uniform, neighbors: 28, metric: euclidean
DT	0.63	0.58	min samples leaf: 200, max depth: 500, criterion: entropy
RF	0.57	0.44	800 estimators, max depth: 10, criterion: entropy
MLP	0.61	0.57	sgd solver, adaptive learning rate, alpha: 0.05, one hidden layer with 20 neurons, activation function: tanh

First we observe that our benchmark baseline model achieves an accuracy of 50% on the test set. This result is not unexpected, given that our dataset is balanced, and the benchmark model always predicts the majority class observed in the training data.

Consequently, as observed for SVM, kNN, and RF, their performance on the test set is slightly below the baseline. This indicates that these models struggle to generalize well to unseen participants, and their increased complexity does not yield any significant improvement. On the other hand, both the MLP classifier and the DT outperform the baseline, although the difference is only marginal. Among the models, the Decision Tree performs the best, achieving an accuracy of 58% on the test set consisting of previously unseen participants.

Examining the test sets' classification report in Table 5 of the best performing model we can see that, for the Stress condition, 71% of instances classified as Stress were indeed participants in the stress condition. However only 27% of the actual participants in the stress condition were correctly classified. For Neutral condition 55% of instances classified as Neutral were indeed participants in the neutral condition. However 89% of the actual participants in the neutral condition are correctly classified. F1-score, a harmonic mean of both precision and recall is 68% for Neutral and 39% for Stress.

Table 5. Classification report of DT for test set

Label	Precision	Recall	F1 score
Neutral	0.55	0.89	0.68
Stress	0.71	0.27	0.39

To further investigate the performance, we can look at the learning curve for the best model. This graph shows how the accuracy of the model evolves as the number of participants, and thus the amount of data, increases. Such a plot can give be informative of if the amount of data is sufficient to successfully generate predictions.

We examined the difference in performance between the SWELL random splitting strategy and our group-wise splitting strategy by comparing the Learning Curves. The results are depicted in Figure 6 on the right and left side respectively. Indeed the performance with the random splitting strategy is increased to those ranges reported in [3].

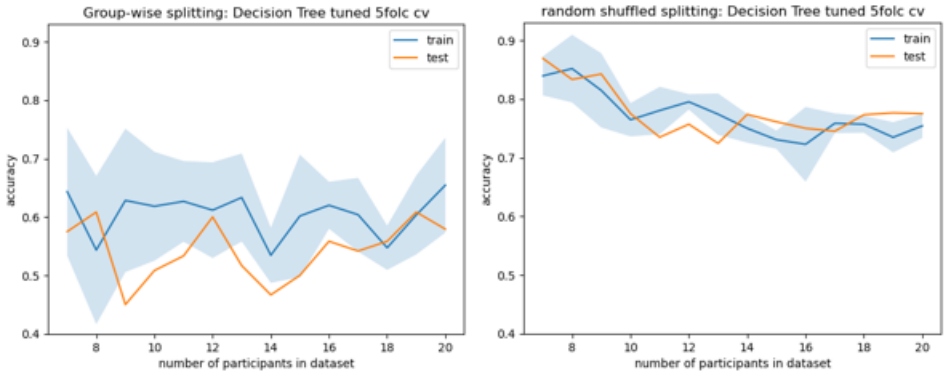

Fig. 6. Learning curve of best model (DT) on train and test set with current, group-wise splitting between train and test (left) and random splitting used in [3] (right). Shaded region represents the standard deviation in the 5-fold CV

Subsequently, we proceeded to create personalized models, where a separate DT model was built for each participant. The results of these personalized models are presented in Figure 7.

It can be observed that all models perform relatively well, with test accuracy ranging from 58% for participant 10 to 100% for several participants, which can be considered good.

4.4 Feature Importance

Finally we proceeded to get the feature importance of each model, in order to examine which features have the most predictive power for accurately detecting stressful situations at work.

The importance of each model/participant over the different features can be seen in Figure 8. Since we have 20 models we pooled the relative importance by summing each individual importance for each feature, and normalising them

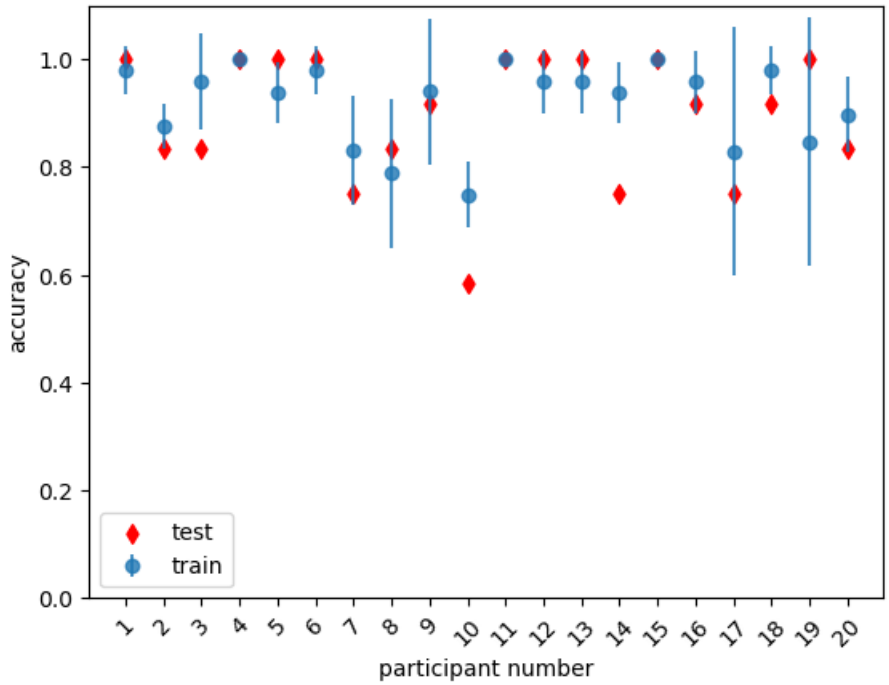

Fig. 7. Performance of personalised DT models on training and testing set, per participant. Error bars represent standard deviation of training accuracy over the 5-fold CV

as described in the methodology. The pooled feature importance over all models/participants is presented in Figure 9

It is evident from both plots that the head orientation features on the x and z direction (left-right and towards-away from screen respectively) show high importance for the majority of personalised models and they also have the highest pooled predictive power. These head orientation features seem to be the most important ones, followed by the "gaze down" and "gaze right" features. Furthermore the typing behavior features "Left click rate", "mouse move distance" and "mouse move duration" also hold predictive power over a number of participant models. Finally we can observe that some facial features are important for some participants, for example the Facial AU "Brow Lowerer" seems to be important only for participant 2.

5 Discussion

In this project we investigated if we can predict stressful working conditions of office workers from unobtrusive sensors available on their work computers. We addressed three distinct aspects of the question through three sub-questions.

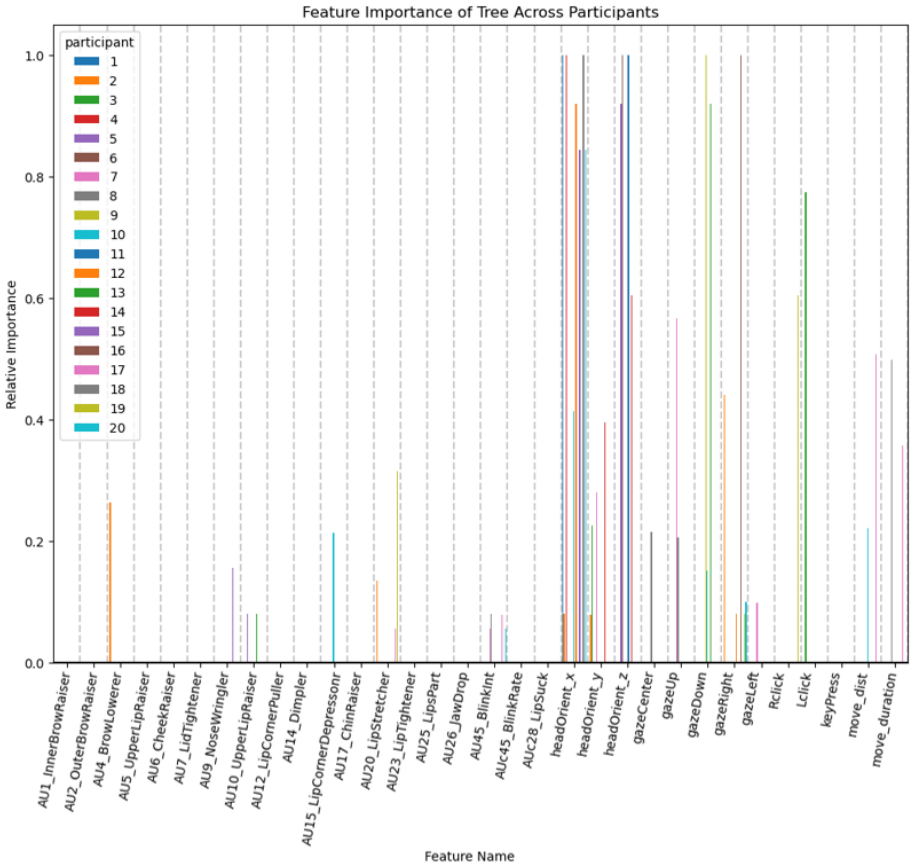

Fig. 8. Feature Importance Across all Participant Models

Firstly, we investigated whether we can induce work-related stress in a controlled experimental setting. The participants' self reported scores on the questionnaires (Figure 4) indicate that we effectively manipulated the condition, as they report experiencing significantly more stress and time pressure on average under the stress condition. However, since questionnaires are subjective and demand the full attention of the user, there is a debate on whether subjective ratings can accurately capture the internal states of an individual [3]. For example in [20] they find that people can mis-remember the "emotional tone" of even the previous work day, hinting that we may not always be fully aware of our actual internal stress levels, and methods like self-report questionnaires could occasionally yield incorrect stress levels.

Consecutively we explored which ML approach is the most suitable for modeling work related stress. The classification results presented in Table 4 demonstrate modest performance compared to the related work presented in this project, with most models performing comparable to the baseline. Among the models

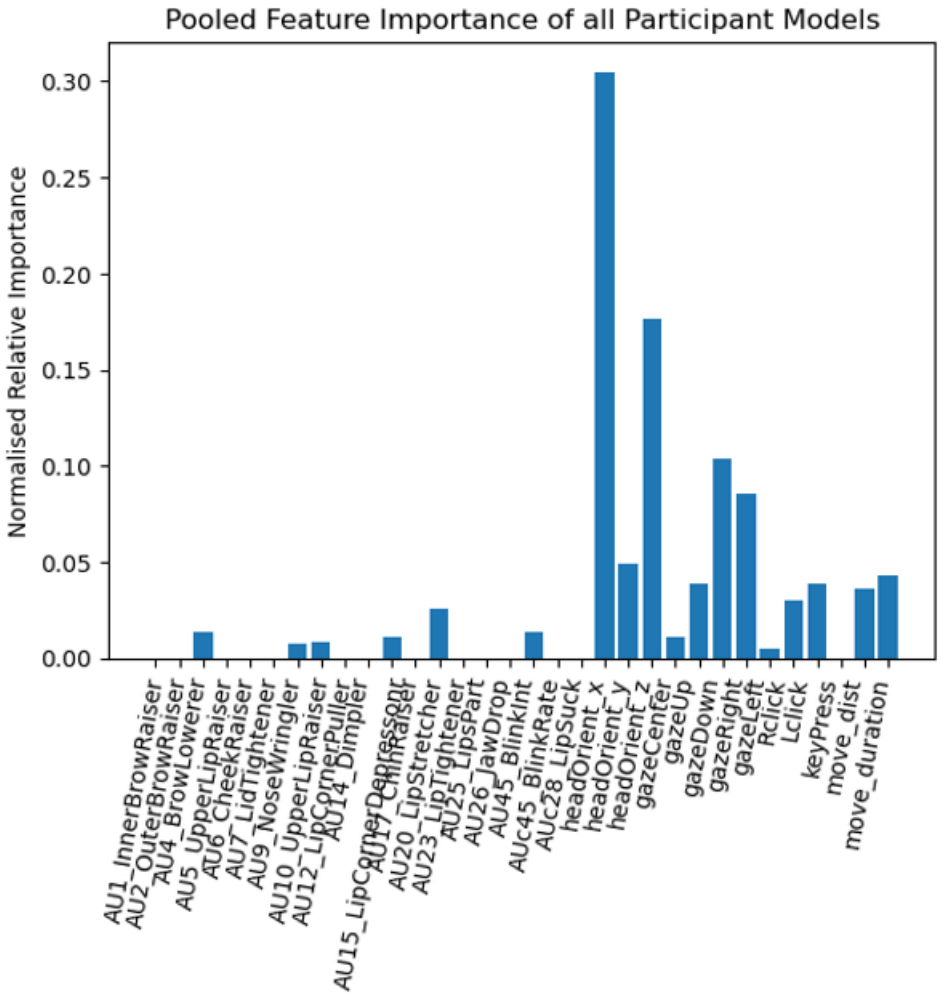

Fig. 9. Pooled Feature Importance of all Participant Models

evaluated, the Decision Tree achieved the highest accuracy of 58% on the test set. It seems that generalizing over new participants is challenging for all models tested here. Moreover, as we were aiming to validate and build upon the findings of [3], we examined the learning curve of this best-performing DT model to determine whether the accuracy had reached a plateau with the available data. Specifically we compared the learning curve of the previously identified best-performing DT model, between the two splitting strategies: random splitting as performed in the SWELL study and group-wise splitting as we were aiming for a model capable of generalizing over unseen participants. It is evident that the splitting strategy greatly impacts on the observed models performance. Splitting

the data randomly resulted in higher accuracy on both the training and testing sets compared to splitting them based on groups. This discrepancy may be attributed to the similarity between the test instances and the training instances, as they consist of different timepoints from the same participants. Alternatively it could also be attributed to the difference in predictive power of the different features per person. However, in both splitting scenarios, an increase in the number of participants led to a slight decrease in performance. This observation suggests that the model performs better with fewer participants, regardless of the splitting strategy. Consequently, this finding may indicate that personalized models, tailored to each individual participant, would be more suitable for this particular task. Indeed, when creating one DT model for each participant, an increase in performance in terms of accuracy, was observed, as depicted in Figure 7. Remarkably, 100% accuracy on the test data was achieved for several participants. This is a strong indication that for the subjective task of classifying stress levels during office work, training a model on an individual basis yields significant benefits. Each person exhibits distinct subtle cues in their typing and facial behavior that reveal their unique stress response.

This conclusion is further supported by the feature importance plot depicted in Figure 8 which encompasses all participant models: facial features such as "Lip Corner Depressor" and "Nose Wrinkler" are important for only one participant each, thereby reinforcing the notion that different individuals exhibit varied but distinct stress cues. Conversely, "Lip Stretcher" and "Upper Lip Raiser" hold importance for three and two participants respectively, suggesting that different groups of people may display similar stress responses. So we have also addressed the last sub-question, namely which features have the most predictive power when predicting stresful situations at work. Overall head orientation features emerge as the most crucial, followed by gaze direction features and some typing behavior features. This observation aligns with the notion that all these aspects could intertwine to form a unique stress response for each individual.

This finding highlights the importance of tailoring models to the specific characteristics of each person when predicting stress levels during office work. Such an approach aligns well with existing research. For instance, studies conducted in [37] and [38] demonstrate that personalized models improve the accuracy of binary stress prediction compared to general models.

6 Limitations

In this paragraph, we will discuss several assumptions that formed the basis of our current work: One assumption we made is that a granularity of 1 minute would be sufficient to predict stress levels. However, due to time and resource constraints, we did not systematically investigate the impact of different granularities on classification accuracy. It is possible that a more fine grained granularity could have captured more detailed changes in facial expressions, typing dynamics and gaze orientation, potentially resulting in the loss of information when aggregating data at 1 minute intervals.

Moreover, it is possible that the classifier is not specifically detecting the working condition of stress or non-stress, but rather the nature of the work task itself. This is because participants were informed that they would only need to give a presentation during the stress block, while the neutral condition involved only report writing. As a result, the typing and mouse dynamics may have been influenced by the task requirements rather than the participants' internal stress levels. Furthermore, the specific stressor we employed, time pressure, can have distinct effects on participants' behavior, such as an increased work pace. Additionally, the behavioral cues of stress may be closely intertwined with each individual's unique work approach and style. These factors may contribute to variations in the observed patterns and highlight the complex nature of stress assessment in office work environments.

Furthermore we used the experimental condition as ground truth for every separate minute of the respective condition. However, it is worth noting that stress is a dynamic and evolving experience that takes time to unfold. Therefore, participants may not be experiencing stress throughout the entire duration of the stress block, as indicated by the ground truth labels. This observation is further supported by the analysis of the confusion matrix. The lower precision for the Neutral label suggests that there are more instances where stress is being misclassified as Neutral (False Positives). Additionally, the low recall of 0.27 for the Stress label indicates that there are a significant number of true stress-labeled instances that are being missed (False Negatives). However, it is important to consider that the classifier may indeed be successfully capturing the distinguishing cues between Stress and Neutral conditions present in the data, but the discrepancies in performance metrics could be attributed to the ground truth labels.

Additionally, it is important to note that the personalized models were trained on a limited amount of data, rendering them susceptible to overfitting. Specifically, the training set consisted of 48 rows per participant (representing 80% of the complete 60-row dataset for each participant). However their performance on the never-seen before test set is reassuring.

7 Conclusion and Future Work

In this thesis study we attempted to classify the working condition of office workers exploiting sensors on their work laptop such as video and typing behavior. For this purpose we conducted an experiment attempting to simulate naturalistic office work setting in a controlled lab environment. The results from the questionnaires indicated that, overall, we successfully created a difference in the self-reported experienced stress and time pressure at the end of the condition. However, it remains unclear whether this difference is present throughout the entire duration of the condition. To answer our main research question we gathered video, mouse and typing behavior data and generated facial behavior features as well as features capturing the typing and mouse usage dynamics. We compared different classification approaches, with the Decision Tree exhibiting the best

performance at a 58% accuracy. Moreover we validated and expanded on the findings of [3] by comparing the learning curve of the best-performing model under two different data splitting strategies, highlighting the impact of the splitting strategy on the observed accuracy. Finally we show that personalised models exhibit improved accuracy, underlining the importance of accounting for individual differences in stress responses. Head and gaze orientation features seem to have the most overarching predictive power across all models.

In terms of future work, as this current thesis was part of a bigger project aiming to estimate Heart Rate measures from video data, the inclusion of estimated HRV as a feature for stress detection should also be explored, as it has shown to be a important stress marker [8–10]. Additionally, the questionnaire results revealed that 50% of participants self-reported increased stress during stress block compared to neutral block. In future work, one could select a subgroup of participants comprising of only those who reported higher stress levels during the stress block. By re-analyzing the data with this subgroup, it would be possible to determine if the observed effects become more pronounced and if the conclusions drawn from the current analysis still hold.

In order to validate our findings in a true naturalistic setting, future work should validate our results on real users of the Welpair Assistant app, within their naturalistic work environments. Since personalized models require some user data to generate predictions, one could use the general model as starting point for new users and retrain the personalized model as we gather facial and typing data from them.

Furthermore all ML models tested in this project do not consider the notion of time, and each minute is classified independently. Future work could explore ML models that explicitly take time into account, such as a Recurrent Neural Network (RNN). An advantage of the RNN apart from its capability to incorporate time sequences, is its ability to integrate feedback from the user on their stress levels during fine-tuning, thus addressing the dynamic nature of stress response during stress conditions.

In conclusion this study contributes to the understanding of the task of stress detection in office environments, it highlights the potential of laptop sensors for stress prediction in the workplace and emphasizes the need for personalized approaches.

References

1. Thorsteinsson, Einar B., Rhonda F. Brown, and Carlie Richards. "The relationship between work-stress, psychological stress and staff health and work outcomes in office workers." Psychology 2014 (2014).
2. Jeunet, C., Mühl, C., and Lotte, F. (2014, January). Design and validation of a mental and social stress induction protocol towards load-invariant physiology-based detection. In International Conference on Physiological Computing Systems.
3. Koldijk, S., Neerincx, M. A., and Kraaij, W. (2016). Detecting work stress in offices by combining unobtrusive sensors. IEEE Transactions on affective computing, 9(2), 227-239.

4. Naegelin, M., Weibel, R. P., Kerr, J. I., Schinazi, V. R., La Marca, R., von Wangenheim, F., ... and Ferrario, A. (2023). An interpretable machine learning approach to multimodal stress detection in a simulated office environment. Journal of Biomedical Informatics, 139, 104299.
5. Suni Lopez, F., Condori-Fernandez, N., and Catala, A. (2019). Towards real-time automatic stress detection for office workplaces. In Information Management and Big Data: 5th International Conference, SIMBig 2018, Lima, Peru, September 3–5, 2018, Proceedings 5 (pp. 273-288). Springer International Publishing.
6. Androutsou, T., Angelopoulos, S., Hristoforou, E., Matsopoulos, G. K., and Koutsouris, D. D. (2023). Automated Multimodal Stress Detection in Computer Office Workspace. Electronics, 12(11), 2528.
7. Panicker, S. S., and Gayathri, P. (2019). A survey of machine learning techniques in physiology based mental stress detection systems. Biocybernetics and Biomedical Engineering, 39(2), 444-469.
8. Innes, G., Millar, W., and Valentine, M. (1959). Emotion and Blood-Pressure By. Journal of Mental Science, 105(440), 840-851. doi:10.1192/bjp.105.440.840
9. Jung, Y., Yoon, Y.I. Multi-level assessment model for wellness service based on human mental stress level. Multimed Tools Appl 76, 11305–11317 (2017). https://doi.org/10.1007/s11042-016-3444-9
10. Bernardi, L., Wdowczyk-Szulc, J., Valenti, C., Castoldi, S., Passino, C., Spadacini, G., and Sleight, P. (2000). Effects of controlled breathing, mental activity and mental stress with or without verbalization on heart rate variability. Journal of the American College of Cardiology, 35(6), 1462-1469.
11. Hjortskov, N., Rissén, D., Blangsted, A. K., Fallentin, N., Lundberg, U., and Søgaard, K. (2004). The effect of mental stress on heart rate variability and blood pressure during computer work. European journal of applied physiology, 92, 84-89.
12. Chen, X., Cheng, J., Song, R., Liu, Y., Ward, R., and Wang, Z. J. (2018). Video-based heart rate measurement: Recent advances and future prospects. IEEE Transactions on Instrumentation and Measurement, 68(10), 3600-3615.
13. Wang, W., Den Brinker, A. C., and De Haan, G. (2018). Single-element remote-ppg. IEEE Transactions on Biomedical Engineering, 66(7), 2032-2043.
14. Okada, Y., Yoto, T. Y., Suzuki, T. A., Sakuragawa, S., Sakakibara, H., Shimoi, K., and Sugiura, T. (2013). Wearable ECG recorder with acceleration sensors for monitoring daily stress. J. Med. Biol. Eng, 33(4), 420-426.
15. Liao, W., Zhang, W., Zhu, Z., and Ji, Q. (2005, September). A real-time human stress monitoring system using dynamic bayesian network. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)-workshops (pp. 70-70). IEEE.
16. Zhai, J., and Barreto, A. (2006, August). Stress detection in computer users based on digital signal processing of noninvasive physiological variables. In 2006 international conference of the IEEE engineering in medicine and biology society (pp. 1355-1358). IEEE.
17. Ren, P., Barreto, A., Huang, J., Gao, Y., Ortega, F. R., and Adjouadi, M. (2014). Off-line and on-line stress detection through processing of the pupil diameter signal. Annals of biomedical engineering, 42, 162-176.
18. Arnrich, B., Setz, C., La Marca, R., Tröster, G., and Ehlert, U. (2009). What does your chair know about your stress level?. IEEE Transactions on Information Technology in Biomedicine, 14(2), 207-214.
19. Dedovic, K., Renwick, R., Mahani, N. K., Engert, V., Lupien, S. J., and Pruessner, J. C. (2005). The Montreal Imaging Stress Task: using functional imaging to

investigate the effects of perceiving and processing psychosocial stress in the human brain. Journal of Psychiatry and Neuroscience, 30(5), 319-325.
20. McDuff, D., Karlson, A., Kapoor, A., Roseway, A., and Czerwinski, M. (2012, May). AffectAura: an intelligent system for emotional memory. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 849-858).
21. Dinges, D. F., Rider, R. L., Dorrian, J., McGlinchey, E. L., Rogers, N. L., Cizman, Z., ... and Metaxas, D. N. (2005). Optical computer recognition of facial expressions associated with stress induced by performance demands. Aviation, space, and environmental medicine, 76(6), B172-B182.
22. https://github.com/TadasBaltrusaitis/OpenFace
23. Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M. A., and Kraaij, W. (2014, November). The swell knowledge work dataset for stress and user modeling research. In Proceedings of the 16th international conference on multimodal interaction (pp. 291-298).
24. Scherer, K. R., and Brody, L. R. (1992). Effects of time-pressure on physiological activity and cognitive performance in a complex problem-solving task. Psychophysiology, 29(6), 570-586
25. Boccignone G, Conte D, Cuculo V, D'Amelio A, Grossi G, Lanzarotti R, Mortara E. pyVHR: a Python framework for remote photoplethysmography. PeerJ Comput Sci. 2022 Apr 15
26. Ahmad, N., Szymkowiak, A., and Campbell, P. A. (2013). Keystroke dynamics in the pre-touchscreen era. Frontiers in human neuroscience, 7, 835.
27. Curtin, M., Tappert, C., Villani, M., Ngo, G., Simone, J., Fort, H. S., and Cha, S. (2006). Keystroke biometric recognition on long-text input: A feasibility study. Proc. Int. MultiConf. Engineers and Computer Scientists (IMECS).
28. Epp, C., Lippold, M., and Mandryk, R. L. (2011, May). Identifying emotional states using keystroke dynamics. In Proceedings of the sigchi conference on human factors in computing systems (pp. 715-724).
29. Gomes, M., Oliveira, T., Silva, F., Carneiro, D., and Novais, P. (2014). Establishing the relationship between personality traits and stress in an intelligent environment. In Modern Advances in Applied Intelligence: 27th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2014, Kaohsiung, Taiwan, June 3-6, 2014, Proceedings, Part II 27 (pp. 378-387). Springer International Publishing.
30. Alberdi, A., Aztiria, A., and Basarab, A. (2016). Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review. Journal of biomedical informatics, 59, 49-75
31. Vizer, L. M., Zhou, L., and Sears, A. (2009). Automated stress detection using keystroke and linguistic features: An exploratory study. International Journal of Human-Computer Studies, 67(10), 870-886.
32. Epp, C., Lippold, M., and Mandryk, R. L. (2011, May). Identifying emotional states using keystroke dynamics. In Proceedings of the sigchi conference on human factors in computing systems (pp. 715-724).
33. Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L. P. (2018, May). Openface 2.0: Facial behavior analysis toolkit. In 2018 13th IEEE international conference on automatic face and gesture recognition (FG 2018) (pp. 59-66). IEEE.
34. https://github.com/TadasBaltrusaitis/OpenFace/wiki/Action-Units
35. Kalman, R.E. (1960): A new approach to linear filtering and prediction problems. J. Basic Eng. 82(1), 35–45
36. Hoogendoorn, M., and Funk, B. (2018). Machine learning for the quantified self. On the art of learning from sensory data.

37. Tazarv, A., Labbaf, S., Reich, S. M., Dutt, N., Rahmani, A. M., and Levorato, M. (2021, November). Personalized stress monitoring using wearable sensors in everyday settings. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 7332-7335). IEEE.
38. Tervonen, J., Puttonen, S., Sillanpää, M. J., Hopsu, L., Homorodi, Z., Keränen, J., ... and Mäntyjärvi, J. (2020). Personalized mental stress detection with self-organizing map: From laboratory to the field. Computers in Biology and Medicine, 124, 103935.